ECOGRAPHY

Research

Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes

Patrice Descombes, Lorenz Walthert, Andri Baltensweiler, Reto Giulio Meuli, Dirk N. Karger, Christian Ginzler, Damaris Zurell and Niklaus E. Zimmermann

P. Descombes (https://orcid.org/0000-0002-3760-9907) ✉ (patrice.descombes@wsl.ch), L. Walthert (https://orcid.org/0000-0002-1790-8563), A. Baltensweiler, D. N. Karger (https://orcid.org/0000-0001-7770-6229), C. Ginzler, D. Zurell (https://orcid.org/0000-0002-4628-3558) and N. E. Zimmermann (https://orcid.org/0000-0003-3099-9604), Swiss Federal Research Inst. WSL, Birmensdorf, Switzerland. DZ also at: Humboldt-Univ. zu Berlin, Berlin, Germany. – R. G. Meuli, Agroscope, Swiss Soil Monitoring Network NABO, Zürich, Switzerland.

Ecography
43: 1–16, 2020
doi: 10.1111/ecog.05117

Subject Editor: Tim Newbold
Editor-in-Chief: Miguel Araújo
Accepted 15 June 2020

Ecologically meaningful predictors are often neglected in plant distribution studies, resulting in incomplete niche quantification and low predictive power of species distribution models (SDMs). Because environmental data are rare and expensive to collect, and because their relationship with local climatic and topographic conditions are complex, mapping them over large geographic extents and at high spatial resolution remains a major challenge.

Here, we propose to derive environmental data layers by mapping ecological indicator values in space. We combined ~6 million plant occurrences with expert-based plant ecological indicator values (EIVs) of 3600 species in Switzerland. EIVs representing local soil properties (pH, moisture, moisture variability, aeration, humus and nutrients) and climatic conditions (continentality, light) were modelled at 93 m spatial resolution with the Random Forest algorithm and 16 predictors representing mesoclimate, land use, topography and geology. Models were evaluated and predictions of EIVs were compared with soil inventory data. We mapped each EIV separately and evaluated EIV importance in explaining the distribution of 500 plant species using SDMs with a set of 30 environmental predictors. Finally, we tested how they improve an ensemble of SDMs compared to a standard set of predictors for ca 60 plant species.

All EIV models showed excellent performance ($|r| > 0.9$) and predictions were correlated reasonably ($|r| > 0.4$) to soil properties measured in the field. Resulting EIV maps were among the most important predictors in SDMs. Also, in ensemble SDMs overall predictive performance increased, mainly through improved model specificity reducing species range overestimation.

Combining large citizen science databases to expert-based EIVs is a powerful and cost–effective approach for generalizing local edaphic and climatic conditions over large areas. Producing ecologically meaningful predictors is a first step for generating better predictions of species distribution which is of main importance for decision makers in conservation and environmental management projects.

Keywords: citizen science, ecological indicator values, Ellenberg, high resolution, humidity, Landolt, pH, soil, species distribution models, Switzerland, wetness

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Predicting the potential distribution of plants has become an important approach in conservation and biodiversity assessments (Guisan and Thuiller 2005, Guisan et al. 2013). So-called species distribution models (SDMs; Guisan and Zimmermann 2000, Guisan et al. 2017), which relate species occurrences or abundances to spatially explicit ecological variables, allow predicting the occurrence probability of a species at a given location across the landscape. The quality of such predictions depends on the input predictors, which should ideally reflect physiological constraints of a species (Guisan and Thuiller 2005, Soberón 2007, Thuiller 2013). So far, climate predictors, such as minimum temperature or growing degree-days, are widely used in plant distribution modelling, likely because of their high availability (Thuiller 2013), accuracy and relevance (Scherrer and Guisan 2019) as well as known direct effects on plant physiology (Woodward 1987, Körner 2003). Similarly, topographic predictors (from digital elevation models) are often used to approximate potential light availability (e.g. solar radiation, aspect) and potential moisture (e.g. topographic wetness index, curvature, slope) in plant distribution models (Zimmermann and Kienast 1999, Thuiller 2013). However, proxies are only surrogates of ecological parameters and may only partially capture the factors relevant for the species' ecological niche. There is a need for more complete and direct predictors enabling to capture a larger spectrum of the species' ecological niche conditions (Austin and Meyers 1996).

Beyond topo-climatic predictors, soil properties strongly affect plant growth and distribution (Elmendorf and Moore 2008, Dubuis et al. 2013). For instance, soils play a main role in driving the distribution of tree species in temperate forests (Walthert and Meier 2017) and ferns in Amazonian forests (Figueiredo et al. 2018). The inclusion of edaphic variables significantly improved the quality of predictions for single plant species as shown for *Acer campestre* L. in France (Coudun et al. 2006). Proxies for soil nutrients, such as soil C:N, were important for modelling the distribution of *Vaccinium myrtillus* L. (Coudun and Gégout 2007), or tree species in the Vosges Mountains (Pinto and Gégout 2005). Geochemical variables (pH and inorganic carbon) and water drainage indicator (volumetric soil water content) improved predictions of alpine plant species (Buri et al. 2020). Similarly, soil pH was the second most important predictor, after temperature, in describing the distribution of 154 alpine plant species in SDMs (Buri et al. 2017). Therefore, considering only climate predictors might lead to incomplete niche quantification of a given species. At the same time, ecophysiologically meaningful variables such as microclimate and soil properties are rarely available (Mod et al. 2016). There is thus a strong need to generate more ecologically meaningful predictors that better reflect local edaphic and climatic conditions and allow better fine scale predictions of species distributions (Mod et al. 2016, Scherrer and Guisan 2019).

Studies investigating site properties at high spatial resolutions (i.e. < 100 m) mostly use parameters measured directly in the field (e.g. pH, moisture, nitrogen). Only few studies have tried to generalize these over large geographic extents (Piedallu et al. 2011, Buri et al. 2017, Hengl et al. 2017), mostly by spatial interpolation (Schloeder et al. 2001), statistical modelling (Häring et al. 2013, Buri et al. 2017) or machine-learning techniques (Heung et al. 2016). In recent decades, great progress has been made to produce digital soil maps also known as digital soil mapping (DSM; Padarian et al. 2019). In DSM, an empirical quantitative relationship is established between soil properties and their spatially implicit soil forming factors such as geology, climate or topography. The methods to relate environmental data and soil properties are manifold and comprise ensemble models of (geo-) statistics, machine- or deep learning (Heung et al. 2016, Nussbaum et al. 2018, Padarian et al. 2019). However, modelling site properties at high spatial resolution remains challenging for two reasons. First, the acquisition of high-resolution climate properties (e.g. local climatic conditions) or soil properties derived in the laboratory from soil samples (e.g. soil pH, soil nutrients) is both time consuming and costly and therefore often leads to sparse datasets (Grunwald et al. 2011, Häring et al. 2013, Carter et al. 2015, Mod et al. 2016). Second, local conditions such as soil moisture and temperature can vary at a very fine spatial scale and are influenced by plants (Aalto et al. 2013, Zellweger et al. 2020). Plants can strongly influence abiotic local conditions and, thus, site properties cannot be accurately predicted without considering the vegetation composition (Aalto et al. 2013). Aalto et al. (2013) found that the inclusion of vegetation variables (i.e. biomass, vegetation volume, lichen and moss cover) strongly improved their models, although local topography and soil properties were the most influential predictors. Similarly, plants have been shown to affect microclimate through temperature buffering (Lenoir et al. 2013), and soil biogeochemistry (i.e. soil acidity and fertility) through variation in the chemistry and quantity of their litterfall (Reich et al. 2005). Similarly, *Sphagnum* species can strongly influence soil moisture and soil pH in peatlands through the production and accumulation of decay-resistant litter (Rydin et al. 2006). Therefore, given the importance of vegetation for modelling local edaphic and climatic conditions, collecting and using large amounts of vegetation data to infer local site properties might provide new insights into generating accurate and ecological meaningful predictors of plants.

Expert-based ecological indicator values (EIVs) are often used in vegetation assessments to provide information on the local abiotic environment (Diekmann 1995, 2003, Wohlgemuth et al. 1999, Gégout et al. 2003, Smart et al. 2003, Wamelink et al. 2005, Häring et al. 2013). In central Europe, local climate and soil properties of plots are often characterized by Ellenberg indicator values averaged among all species found in these plots (Ellenberg et al. 1992), or by Landolt indicator values in Switzerland and the European

Alps (Landolt et al. 2010). These two EIV systems represent subjective categorical assessments of the response of species to different environmental conditions, such as soil moisture, soil acidity (pH), soil nutrients (mainly nitrogen), light, temperature and continentality. While Ellenberg et al. (1992) derived those metrics for approximately 2700 species in Central Europe, Landolt et al. (2010) assessed EIVs for almost all plant species occurring in Switzerland (ca 3600 species). EIVs are often used as bioindicators of the site characteristics when plant inventories are available (Ter Braak and Barendregt 1986). For instance, Scherrer and Guisan (2019) used EIVs averaged at the plot level and found that ecological conditions reflecting light availability and soil conditions are, in complement to temperature, very important and should be included in SDMs. While predictions of EIVs at the plot scale are common practice, assessing the potential of combining large amounts of plant occurrences and EIVs to predict local soil and climate properties (derived from EIVs) over large areas at high spatial resolution (i.e. < 100 m) has only received limited consideration so far (Häring et al. 2013).

Here, we propose to combine plant occurrence databases with EIVs to predict local edaphic and climatic conditions over large geographic extents and at high spatial resolution (93 m). First, using two independent datasets of combined soil and plant inventories distributed in Switzerland, we investigated how averaged site EIVs across inventoried species relate to local values of measured soil properties (i.e. pH, Ca, total nitrogen content, organic carbon content, C:N, sum of basic cations, hydromorphy, modelled drought index). Then, we combined citizen science-based data of plant occurrences from Switzerland to expert-based EIVs for ca 3600 plant species and averaged EIVs at the grid cell level (93 m) over the entire landscape. We related site-averaged EIVs to topographic, mesoclimatic, landuse and geological predictors at more than 70 000 sites using Random Forest models and mapped them across Switzerland at a 93 m resolution. For comparison, we also modelled and mapped soil properties (pH, total nitrogen content, organic carbon content, C:N) with the same methodology. Finally, we investigated how the generated variables improved predictions of individual plant species in Switzerland using SDMs.

Material and methods

Study area

The study area encompasses Switzerland, a European country presenting a temperate climate and a complex topography, with an elevation gradient ranging from 190 m to ca 4500 m a.s.l., with diverse edaphic and climatic conditions from moist to dry environments and calcareous to acidic soils.

Plant ecological indicator values (EIVs)

We obtained plant EIVs for 3599 species from Flora Indicativa (Landolt et al. 2010). We retained 8 different EIVs to characterize the local edaphic and climatic conditions: soil pH (EIV-R), soil nutrients (EIV-N, i.e. mainly nitrogen), soil moisture (EIV-F), soil moisture variability (EIV-W), soil aeration (EIV-D), soil humus (EIV-H), continentality (EIV-K) and light (EIV-L). EIVs reflecting tolerance to salt (EIV-S) and heavy metal (EIV-M) were not considered. The temperature EIV (EIV-T) was also not considered, as it highly correlates with temperature from climate models (Scherrer and Guisan 2019). All 8 EIVs are ordinal variables consisting of 3–9 classes each (see Table 1 for a summarized description). Because EIV-R classes in the Flora Indicativa represent distinct pH ranges (e.g. pH 2.5–5.5 for class 1), we retained the median pH value for each class (classes 1–5 correspond to pH values of 4, 5, 6, 7 and 7.5, respectively) to make this parameter more comparable to in-situ pH values. Plant species presenting undefined EIV or broad ecological preferences ('–' or 'x' in the Flora Indicativa) were not considered in further analyses (representing less than 0.84% of the species).

Plant distribution data

We used distribution data from three different datasets: 1) citizen–science based plant occurrence data, 2) a vegetation plot inventory and 3) a forest plant inventory.

First, the citizen science data containing ~6.7 million records of plant occurrences was obtained from the National

Table 1. Description of the ecological indicator values (EIVs) used in this study. See Landolt et al. (2010) for a detailed description of the EIVs.

EIVs	Description	Ordinal classes	Description of classes
EIV-R	Soil pH	1, 2, 3, 4, 5	Gradient from acidic soils (1) to carbonate containing alkaline soils (5)
EIV-N	Soil nutrients	1, 2, 3, 4, 5	Gradient from nutrient-poor soils (1) to nutrient-rich soils (5), mainly nitrogen
EIV-F	Soil moisture	1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5	Gradient from very dry soils (1) to plants growing in water (5)
EIV-W	Soil moisture variability	1, 2, 3	Gradient from low intraannual variability in soil moisture (1) to soils with a high intraannual variability in soil moisture (3)
EIV-D	Soil aeration	1, 3, 5	Gradient from waterlogged/low aerated soils (1) to soils rich in rocks or sand with larger distance to the water table (5)
EIV-H	Soil humus	1, 3, 5	Gradient from humus-poor soils (1) to humus-rich soils (5)
EIV-K	Continentality	1, 2, 3, 4, 5	Gradient from atlantic climate (1; high mean air humidity, low variations in temperature and relatively mild winters) to continental climate (5; low mean air humidity, high variations in temperature and cold winters)
EIV-L	Light	1, 2, 3, 4, 5	Gradient from shaded (1) to sunny areas (5)

Data and Information Center on the Swiss Flora (Info Flora: <www.infoflora.ch/>; data extracted in April 2019). We retained only geographically valid occurrences with a coordinate precision <= 100 m, corresponding to ~5.1 million occurrences of 3981 species collected between years 1565–2019 (< 1% records before 1956, 10th percentile = 1996, 90th percentile = 2017). This database contains also information on the type of observation. An observation can be opportunistic or part of an exhaustive plant inventory. These data serve as input to modelling EIVs and to modelling plant species distributions.

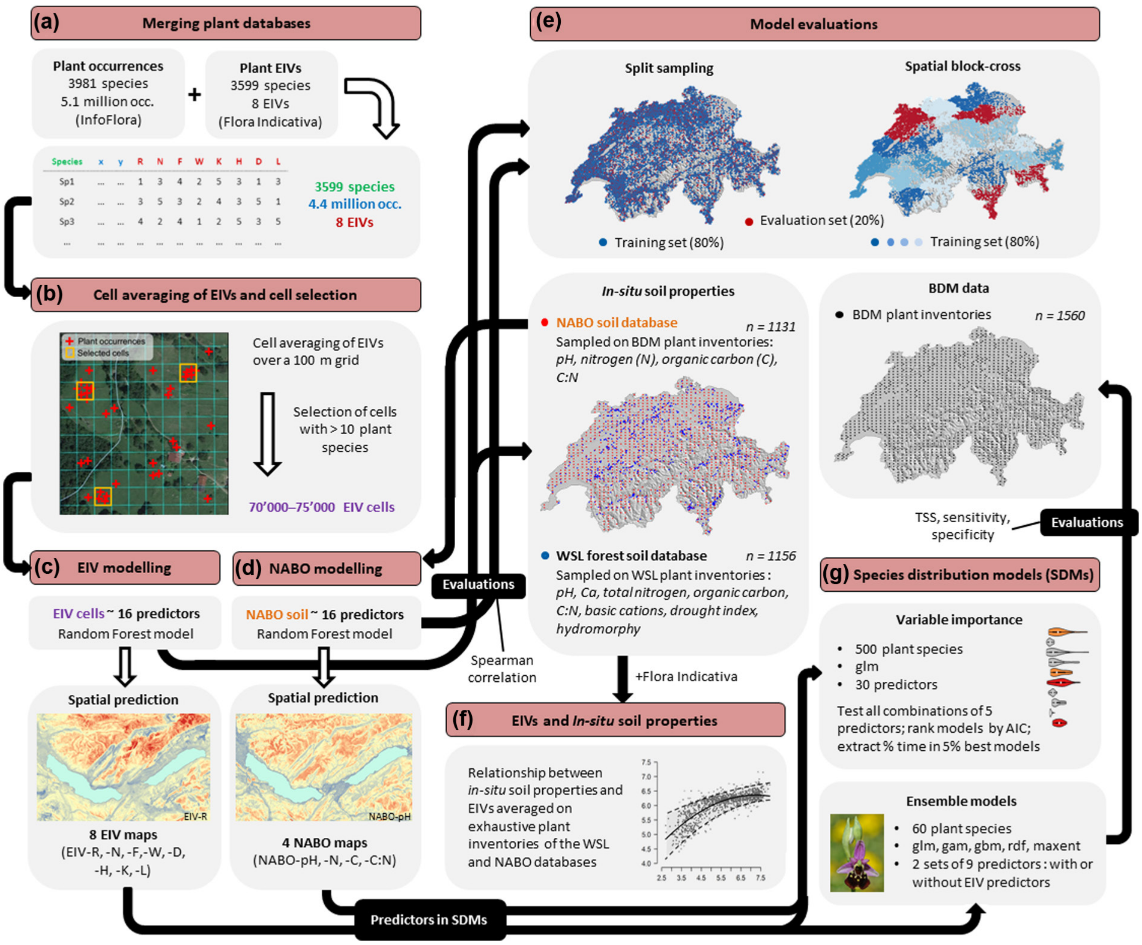
Second, vegetation plot inventories were performed in 5-yr intervals between 2000 and 2018 on 1560 vegetation plots of 10 m^2 at intersections of a 1 km grid placed over Switzerland within the Swiss Biodiversity Monitoring survey (BDM indicator Z9 'species diversity in habitats', <www.biodiversitymonitoring.ch>; Fig. 1). The BDM plant inventory dataset includes vegetation plots spanning different types of habitats from low to high elevation. We considered all recorded plants across all replicated inventories for each plot location. The BDM dataset was used as an independent dataset for validating the relationship between EIVs and soil properties, and for validating final SDMs.

Third, forest plot inventories were performed between 1938 and 2014 (95% after 1993) on 1156 expert-selected forest vegetation plots of 100–500 m^2 (avg. 200 m^2) across Switzerland (Fig. 1), conducted by about 50 botanists at sites where soil inventories (see description below) were also performed by the Swiss Federal Inst. for Forest, Snow and Landscape Research (WSL). See Walthert and Meier (2017) for further details on the sampling methodology. These data were used as an independent dataset for validating the relationship between EIVs and soil properties.

Soil databases

Datasets on soil properties were derived from forest soil profiles sampled by WSL where forest plant inventories were also performed (hereafter called WSL soil database) and from a topsoil inventory performed on a subset of the BDM vegetation plots within the Swiss Soil Monitoring Network NABO (hereafter called NABO soil database).

The WSL soil database contains 1156 forest soil profiles that are, on average, 1.2 m deep and sampled across Switzerland between 1938 and 2014 (95% after 1993). Soil sampling as well as physical and chemical soil analyses were


(a) Merging plant databases
Plant occurrences
3981 species
5.1 million occ.
(InfoFlora)
+
Plant EIVs
3599 species
8 EIVs
(Flora Indicativa)

Species	x	y	R	N	F	W	K	H	D	L
Sp1	...	...	1	3	4	2	5	3	1	3
Sp2	...	...	3	5	3	2	4	3	5	1
Sp3	...	...	4	2	4	1	2	5	3	5
...	...	...	...	...	...	...	...	...	...	...

3599 species
4.4 million occ.
8 EIVs
(b) Cell averaging of EIVs and cell selection
Plant occurrences
Selected cells
Cell averaging of EIVs over a 100 m grid
Selection of cells with > 10 plant species
70'000–75'000 EIV cells
(c) EIV modelling
EIV cells ~ 16 predictors
Random Forest model
Spatial prediction
EIV-R
8 EIV maps
(EIV-R, -N, -F, -W, -D, -H, -K, -L)
(d) NABO modelling
NABO soil ~ 16 predictors
Random Forest model
Spatial prediction
NABO-pH
4 NABO maps
(NABO-pH, -N, -C, -C:N)
Evaluations
Spearman correlation
Predictors in SDMs
(e) Model evaluations
Split sampling
Spatial block-cross
Evaluation set (20%)
Training set (80%)
Training set (80%)
In-situ soil properties
NABO soil database
n = 1131
Sampled on BDM plant inventories:
pH, nitrogen (N), organic carbon (C), C:N
WSL forest soil database
n = 1156
Sampled on WSL plant inventories:
pH, Ca, total nitrogen, organic carbon, C:N, basic cations, drought index, hydromorphy
+Flora Indicativa
BDM data
BDM plant inventories
n = 1560
TSS, sensitivity, specificity
Evaluations
(f) EIVs and In-situ soil properties
Relationship between in-situ soil properties and EIVs averaged on exhaustive plant inventories of the WSL and NABO databases
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
2.5
3.5
4.5
5.5
6.5
7.5
(g) Species distribution models (SDMs)
Variable importance
500 plant species
glm
30 predictors
Test all combinations of 5 predictors; rank models by AIC; extract % time in 5% best models
Ensemble models
60 plant species
glm, gam, gbm, rdf, maxent
2 sets of 9 predictors: with or without EIV predictors

Figure 1. Overview of the main analyses, variables and databases used in the present study.

performed in each soil horizon of the soil profiles and are described in detail in Walthert et al. (2013) and in Walthert and Meier (2017). Soil pH (pH $CaCl_2$), total nitrogen content ($g kg^{-1}$), organic carbon content ($g kg^{-1}$), organic carbon to total nitrogen ratio (C:N), Ca content ($mmolc kg^{-1}$) and sum of Ca, Mg, K and Na (basic cations; $mmolc kg^{-1}$) were calculated as the average of all horizons between 0 and 25 cm soil depth including eventual organic topsoil horizons. For each soil profile, we estimated a drought index, a proxy of soil moisture availability and the soil hydromorphy, a proxy of soil aeration or periodic oxygen shortage (Supplementary material Appendix 1 Note A1). Note that the WSL soil database contains, in total, 1204 forest soil profiles, but we kept only soil profiles which have both plant inventories and all soil properties available (n = 1156).

The NABO soil database was collected between 2011 and 2015 on 1131 locations selected among the 1560 vegetation plots of the BDM vegetation database (BDM indicator Z9 'species diversity in habitats'). On each plot, 2–4 topsoil samples at levels down to 20 cm (including organic horizons) were collected in each corner around the central circular BDM vegetation plot (Meuli et al. 2017) for a description of the sampling). Soil pH (pH $CaCl_2$), total nitrogen content (N; $g kg^{-1}$), organic carbon content (C; $g kg^{-1}$) and organic carbon to total nitrogen ratio (C:N) were measured on each sample, and averaged for each vegetation plot. Note that some soil properties were not measured on some sites (pH: n = 1130; organic carbon: n = 1103; nitrogen: n = 1114; C:N: n = 1103).

Environmental predictors

Environmental predictors are used for modelling distribution of EIVs across space, and for SDMs. We used a digital elevation model (DEM) at 93 m resolution (Robinson et al. 2014) as a base map to derive different topographic, landuse, geological, remote sensing and climate properties calculated or extracted from various sources (Supplementary material Appendix 1 Table A1) and known to be important for plants (see Mod et al. 2016 and references therein). Predictors included aspect, topographic position index (TPI), topographic roughness index (TRI), topographic wetness index (TWI), distance to waters, distance to buildings, yearly sum of precipitation (Prec year), sum of winter precipitation (Prec 12–2), sum of summer precipitation (Prec 6–8), yearly temperature average (Tave year), winter temperature average (Tave 12–2), summer temperature average (Tave 6–8), growing degree days above 5.56°C (Gdd), yearly atmospheric moisture balance (Mind), yearly sum of global (direct and diffuse) potential solar radiation (Srad year), annual average site water balance (Swb), forest canopy height (Forest height), inner forest density (Forest height Q25), mean summer normalized difference vegetation index (NDVI mean), variability in summer normalized difference vegetation index (NDVI SD), x-coordinate, y-coordinate and bedrock geology. Detailed descriptions of these predictors can be found in Supplementary material Appendix 1 Note A2.

In-situ relationship between site-averaged EIVs of plant inventories and soil properties measured in the field

We analysed the in-situ relationship between site-averaged EIVs of plant inventories and soil properties measured in the field within the NABO and WSL soil databases (Fig. 1f). To do so, we linked the species lists per site to EIVs (Landolt et al. 2010) using the species' taxonomic identity code (SIN, isfs number). We then calculated average EIV values for each site based on the plant inventory by considering only the presence of the species, meaning that we did not account for relative abundances of species. Average plot EIV and in-situ soil measures were compared by using Spearman's rank correlation test and adjusted R-square (R^2) of a linear model including linear and quadratic terms. Values of the coefficient of correlation r (positive or negative) above 0.70 are considered as strong, between 0.30 and 0.69 as moderate and below 0.29 as weak (Fowler et al. 2013). We compared EIV-R to pH (NABO and WSL soil databases) and Ca (WSL), EIV-N to total nitrogen content and C:N (NABO and WSL) and sum of basic cations (WSL), EIV-H to organic carbon content (NABO and WSL), EIV-F to modelled drought index (WSL) and EIV-D to hydromorphy (WSL).

Mapping EIVs

We calibrated eight different EIV models for the different indicators (EIV-R, EIV-N, EIV-F, EIV-W, EIV-D, EIV-H, EIV-K and EIV-L) and then mapped these EIVs across Switzerland (10 152 417 grid cells in total). To do so, we first combined the ~5.1 million InfoFlora plant occurrences to their EIVs (Landolt et al. 2010) using their taxonomic identity code (SIN, isfs number; Fig. 1a). Then, we averaged EIVs in each grid cell at 93 m resolution after removing duplicate observations of species occurring in the same grid cell. To ensure that our averaged cell EIVs are robust and representative of the edaphic and climatic conditions, we retained only cells with at least ten records of different plant species (~ 73 000 cells; Fig. 1b, Supplementary material Appendix 1 Fig. A1). We tested the robustness of species and site selection by testing thresholds of >= 5 and >= 20 species, and by including only observations that are part of exhaustive plant inventories (n = 86 715 plant inventories). In addition, we checked how cell average EIVs are affected if we remove one single species (for 500 species; see Supplementary material Appendix 1 Table A2 for the species list) at a time by using a Spearman's rank correlation test and calculated the number of cell loss (in %) resulting from lower number of cells with at least ten records of species. We then related cell averaged EIVs to a set of climatic, topographic, landscape and geological predictors (Supplementary material Appendix 1 Table A1) and predicted EIVs over Switzerland (Fig. 1c). All selected predictors showed pairwise Pearson correlations $|r| < 0.8$ (Dormann et al. 2013). We used Random Forests (Breiman 2001) with 1000 trees and a node size of 20 to reduce computation time and avoid overfitting. We assessed the contribution (in %) of each predictor in the EIV modelling by permutation importance. For this, we reshuffled randomly each predictor individually (keeping all other predictors unchanged), and measured the decrease in prediction accuracy on the reshuffled data with a Pearson correlation (Strobl et al. 2007). We repeated the reshuffling five times, transformed the average accuracy loss to % contribution, and report this value as a measure of variable importance. Because model extrapolation to novel covariates in space may be found in Switzerland due to the low environmental coverage of the calibration data, we quantified the level of extrapolation outside the univariate range (type 1 novelty) and novel covariate combinations within the univariate range of covariates (type 2 novelty) following Mesgaran et al. (2014). We calculated the percentage of cells affected by both types of extrapolation and determined the most influential covariates responsible for them.

EIV models were evaluated by five-fold split-sampling of the data (training set = 80%, evaluation set = 20%) and by five-fold spatial block cross-validation (Roberts et al. 2017) using five strata assigned across 25 regional clusters in Switzerland (training set = 80%, evaluation set = 20%) and by using Spearman's rank correlation tests (Fig. 1e).

Model predictions were also externally evaluated on the two independent datasets of in-situ soil properties (NABO and WSL soil databases) by calibrating the models without the cell grids containing soil inventories (Fig. 1e). Predictions and in-situ measures were compared by using Spearman's rank correlation tests. We expected to find lower correlations but similar directions of correlations when using model-predicted than when using site-averaged EIVs of plant inventories performed on the soil sampling location. Finally, we used the Swiss federal inventories (polygons) of wetlands (fens and bogs) and the Swiss federal inventories (polygons) of dry meadows and pastures (<www.bafu.admin.ch/>) to validate predictions of EIV-F, EIV-W and EIV-D over Switzerland. We averaged predicted EIV-F, EIV-W and EIV-D, as well as other proxies of moisture variability (distance to water, annual average site water balance and moisture index), for each polygon and compared wetlands and dry grasslands with a Welch two sample t-test. Because wetlands and dry grasslands both occur at different extremes of the moisture gradient, we expect to find non-overlapping mean ± SD of EIV's between wet and dry polygon objects.

Mapping in-situ soil properties

To investigate if EIV maps reveal better predictive performance than mapped in-situ soil properties in SDMs, we also mapped the in-situ soil properties of the NABO soil database (Fig. 1d). In-situ soil properties of the WSL database were not modelled because they only cover forest habitats. We used the same predictors as for mapping the EIVs, Random Forests (Breiman 2001) with 1000 trees and a node size of 1, and predicted the NABO in-situ properties over Switzerland (NABO-pH, NABO-N, NABO-C, NABO-C:N). We quantified the level of extrapolation in the same way as for mapping EIVs. NABO models were evaluated by five-fold split-sampling and by five-fold spatial block cross-validation.

EIV importance in species distribution models

We investigated the importance of 30 predictors (including the generated 8 EIV and 4 NABO maps; Supplementary material Appendix 1 Table A1) in explaining the distribution of 500 plant species in Switzerland by using generalized linear models (McCullagh 1983; Fig. 1g). The 500 species were selected to represent frequent native plant species with at least 250 occurrences across Switzerland and to span different ecological groups (see Supplementary material Appendix 1 Table A2 for the species list). We used geographically valid and precise (<= 100 m) occurrences from the InfoFlora database, which were thinned to 300 m to reduce spatial autocorrelation in model residuals. Species' occurrences were related to 10 000 pseudo-absences that were sampled randomly across Switzerland and with the same spatial bias as observed plant distribution patterns (Phillips et al. 2009). We ran an inverse distance weighted spatial model by using the 'geoIDW' function in the 'gstat' package (Pebesma 2004), and by selecting randomly the same number of pseudo-absences as the number of occurrences of the modelled species. This geographic model informs on the probability of being close to an occurrence in the landscape by means of X and Y coordinates and has the advantage of being independent of a distance threshold. All probabilities < 0.1 were set to 0.1 to have a minimum 10% probability of sampling across the landscape. We then selected 10 000 pseudo-absences over this weighted probability map with a minimal distance of 300 m between pseudo-absences. Then, we ran generalized linear models with binomial distribution with all possible combinations of a maximum of five predictors including quadratic terms using the 'dredge' function in the 'MuMIn' package (Barton 2019), resulting in > 80 000 combinations per species. For computation time reasons, interactions between environmental predictors were not included. Models presenting pairs of highly correlated predictors (Pearson correlations: $|r| > 0.8$) were not included (Dormann et al. 2013). Variable importance was assessed as the percentage of time a variable was present in the 5% best models ranked by their AICs (4042 out of 80 845 models for every species). We compared the importance of each predictor for all species and between ecological groups from Flora Indicativa (Landolt et al. 2010).

Comparing species distribution models with and without EIV predictors

We assessed the potential distribution of 60 plant species at a 93 m spatial resolution in Switzerland by using ensemble SDMs (SDMs following Guisan and Zimmermann 2000) with five algorithms and five pseudo-absence iterations (Fig. 1g). For computational reason, we only selected 60 plant species being a representative set of native plant species observed at least 150 times along the elevation gradient and presenting different ranges of ecological preferences from dry to moist and acidic to alkaline soil conditions (see Supplementary material Appendix 1 Table A3 for the species list). SDMs either included or excluded EIVs. Spatial predictions were validated against BDM data using different performance measures (specificity, sensitivity, true skill statistic TSS; Allouche et al. 2006). Methodological details are provided in Supplementary material Appendix 1 Note A3. All analyses were performed in R ver. 3.5.1 (R Core Team).

Results

In-situ relationship between site-averaged EIVs of plant inventories and soil properties measured in the field

We found strong positive correlations and relationships between EIV-R and soil pH measured in the soil profiles in both independent soil databases (NABO: r = 0.768, n = 1103, R^2 = 0.652; WSL: r = 0.819, n = 1156, R^2 = 0.628; Fig. 2a, d). EIV-R and Ca were moderately positively correlated (WSL: r = 0.705, n = 1156, R^2 = 0.438; Supplementary material Appendix 1 Fig. A2). EIV-N and C:N were moderately and negatively correlated both in the NABO (r = −0.637, n = 1079, R^2 = 0.325; Fig. 2b) and WSL (r = −0.608; n = 1156, R^2 = 0.393; Fig. 2e) soil databases. In contrast, we found only weak correlations and relationships between EIV-N and the total nitrogen content (NABO: r = −0.191, n = 1089, R^2 = 0.038; WSL: r = −0.179, n = 1156, R^2 = 0.034), as well as between EIV-N and the sum of basic cations (WSL: r = 0.163, n = 1156, R^2 = 0.022). We found a moderate positive correlation between EIV-F and the modelled drought index (WSL: r = 0.393, n = 1139, R^2 = 0.307; Fig. 2c), and a moderate negative correlation between EIV-D and hydromorphy (WSL: r = −0.439, n = 1156, R^2 = 0.239, Fig. 2f). We found a weak positive correlation between EIV-H and organic carbon content measured in the WSL soil profiles (WSL: r = 0.182, n = 1156, R^2 = 0.045; Supplementary material Appendix 1 Fig. A2), and a moderate positive correlation in the NABO soil database (NABO: r = 0.449, n = 1079, R^2 = 0.202; Supplementary material Appendix 1 Fig. A2). Correlation coefficients are summarized in Table 2 (column 'r site').

EIV model evaluations

All EIV Random Forest models provided very good test statistics with strong positive correlations when evaluated by split sample testing (Spearman correlation: r > 0.918; Table 2) and block cross-validation (r > 0.904; Table 2). Overall, results were only marginally affected by the choice of the threshold for averaging EIVs on cells (Supplementary

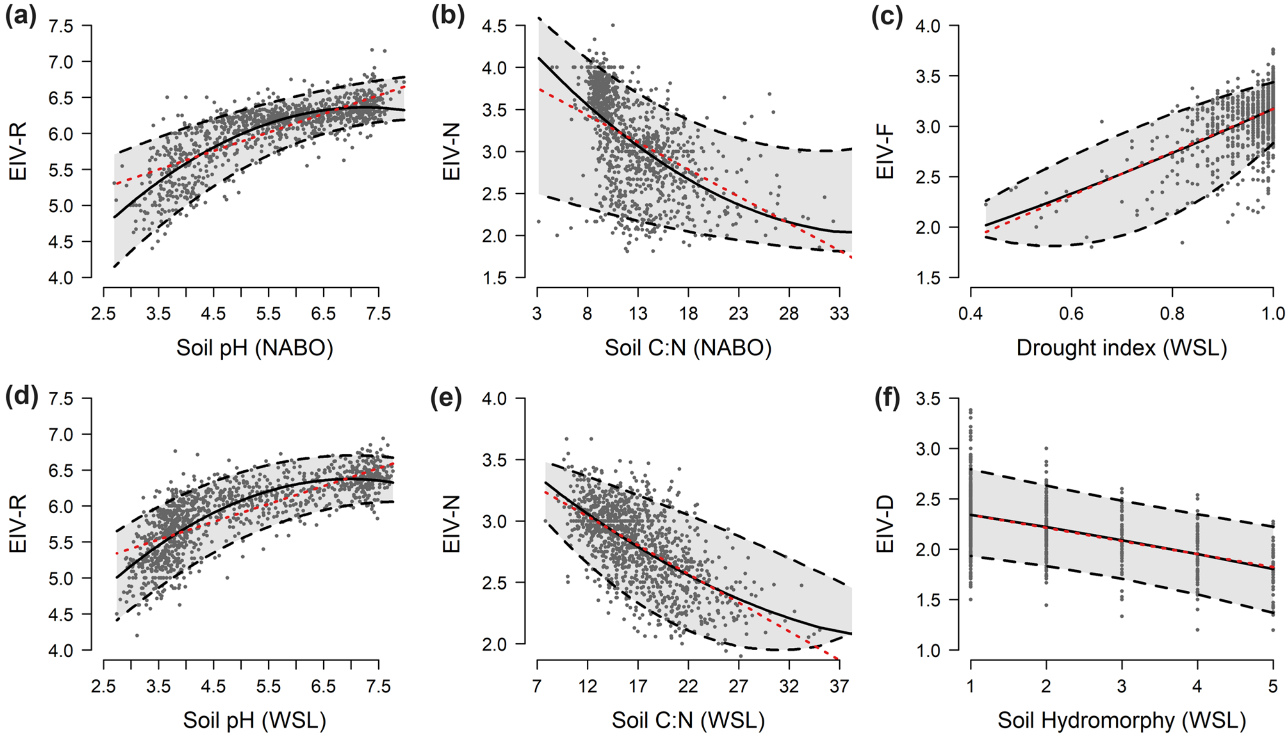


Figure 2. Relationships between site-averaged EIVs of plant inventories (y axis) and in-situ soil properties measured in the field (x axis) within the NABO (a–b) and WSL (c–f) soil databases. Each dot represents a vegetation inventory where EIVs were averaged across all plant species and where in-situ soil properties were measured. The black lines represent the quadratic relationships and the grey areas span from the 5th to the 95th percentiles. The dotted red lines represent the linear relationships. Note that EIV-R classes were converted to the pH ranges described in the Flora Indicativa (Landolt et al. 2010) to make this parameter more comparable to in-situ pH values (see Method section). EIVs are described in Table 1 and correlation coefficients between site-averaged EIVs and in-situ soil properties are summarized in Table 2 (column 'r site').

Table 2. Spearman rank correlations between observed and predicted EIVs based on repeated split-sampling (Split) and five-fold block cross-validation (CV) tests. Models were also evaluated by comparing in-situ soil properties measured on sites from two soil databases (NABO and WSL) to predicted EIVs (r model). Relationships between in-situ soil properties measured in the field (NABO and WSL) and site-averaged EIVs of exhaustive plant inventories at the same sites (r site) are also provided. EIVs are described in Table 1. Correlation coefficients |r| > 0.3 are highlighted in bold.

	Split		CV		WSL soil database			NABO soil database		
	r		r							
EIVs	Mean	SD	Mean	SD	Soil properties	r (model)	r (site)	Soil properties	r (model)	r (site)
EIV-R	**0.924**	0.001	**0.904**	0.000	pH	**0.607**	**0.819**	pH	**0.735**	**0.768**
					Ca	**0.530**	**0.705**	–	–	–
EIV-N	**0.938**	0.001	**0.939**	0.000	Nitrogen	**−0.351**	−0.179	Nitrogen	−0.230	−0.191
					C:N	**−0.348**	**−0.608**	C:N	**−0.589**	**−0.637**
					Basic cations	−0.201	0.163	–	–	–
EIV-F	**0.946**	0.001	**0.941**	0.000	Drought index	**0.411**	**0.393**	–	–	–
EIV-W	**0.924**	0.002	**0.925**	0.000	–	–	–	–	–	–
EIV-D	**0.936**	0.001	**0.939**	0.000	Hydromorphy	**−0.590**	**−0.439**	–	–	–
EIV-H	**0.932**	0.002	**0.909**	0.000	Organic carbon	−0.001	0.182	Organic carbon	**0.418**	**0.449**
EIV-K	**0.931**	0.001	**0.930**	0.000	–	–	–	–	–	–
EIV-L	**0.918**	0.001	**0.926**	0.000	–	–	–	–	–	–

material Appendix 1 Table A4), and all cell averaged EIVs were robust against removing single species from the occurrence database (r > 0.999; cell loss: < 0.071%). The variable importance, measured as contribution of each predictor in the EIV modelling, varied between 0.2 and 40% (Supplementary material Appendix 1 Fig. A3). EIV model extrapolation into novel covariate space was low (type 1 novelty = ~0.9%; type 2 novelty < 0.01%) and mostly concerned high elevation sites presenting temperatures below the range of the calibration data (> 3200 m; Supplementary material Appendix 1 Fig. A4, A5). The generated EIV maps are visualized in Fig. 3, Supplementary material Appendix 1 Fig. A6.

Evaluations of model predictions on the two independent datasets of in-situ soil properties presented moderate to strong positive correlations between EIV-R and pH (WSL: r = 0.607; NABO: r = 0.735), and a moderate positive correlation between EIV-R and Ca (WSL: r = 0.530). Evaluations of model predictions presented a moderate negative correlation between EIV-N and C:N (WSL: r = −0.348; NABO: r = −0.589), while, in contrast, correlations between EIV-N and total nitrogen content (WSL: r = −0.351; NABO: r = −0.230) or sum of basic cations (WSL: r = −0.201) were generally weak. Among soil moisture proxies, evaluations of model predictions presented a moderate positive correlation between EIV-F and modelled drought index (WSL: r = 0.411), and a moderate negative correlation between EIV-D and hydromorphy (WSL: r = −0.590). Evaluations between EIV-H and organic carbon content presented weak to moderate positive correlations (NABO: r = −0.001; WSL: r = 0.418). Correlations between EIV predictions and in-situ soil properties were generally lower (column ‘r model’ in Table 2) than between site-averaged EIVs of plant inventories

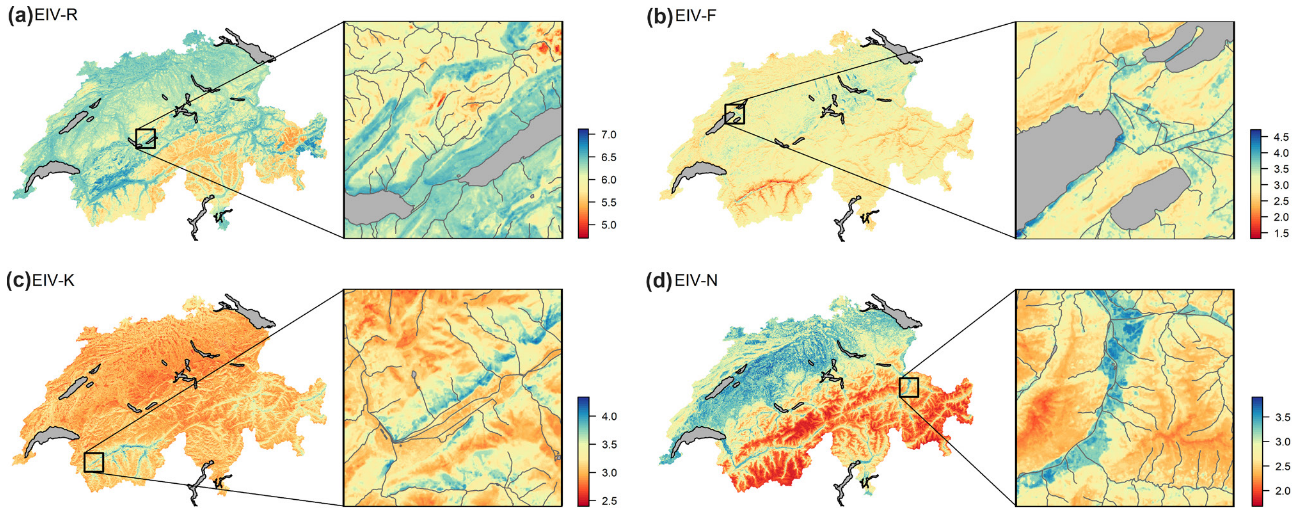


Figure 3. Maps of four EIVs (a–d) predicted across Switzerland. Cell-averaged EIVs were related to a set of climate, landuse, topographic and geological predictors using Random Forest models, and were predicted across the entire landscape. The focal region (box) represents a square of 30 km width, and illustrates predicted EIVs in more detail. Lakes and rivers are mapped in grey. EIVs are described in Table 1. Maps of all EIVs are presented in Supplementary material Appendix 1 Fig. A6.

and soil properties measured in the field (column 'r site' in Table 2), but presented similar directions of correlations. Correlation coefficients are summarized in Table 2.

EIV-F, EIV-W and EIV-D models correctly predicted areas with wetlands (fens and bogs) and dry meadows in Switzerland (Supplementary material Appendix 1 Fig. A7). EIV-F and EIV-W were both significantly higher in wetlands than in dry grasslands with non-overlapping mean ± SD (mean ± SD; EIV-F: moist = 3.37 ± 0.24, dry = 2.63 ± 0.27, t = 153.04, df = 11089, p-value < 0.001; EIV-W: moist = 2.11 ± 0.21, dry = 1.71 ± 0.13, t = 127.11, df = 9942, p-value < 0.001; Supplementary material Appendix 1 Fig. A8). EIV-D was significantly lower in wetlands than in dry grasslands with non-overlapping mean ± SD (mean ± SD; moist = 1.78 ± 0.28, dry = 2.44 ± 0.27, t = −130.05, df = 11341, p-value < 0.001; Supplementary material Appendix 1 Fig. A8). In contrast, proxies of moisture gradients such as distance to water (Water), annual average site water balance (Swb) and moisture index (Mind) also presented significant differences between wetlands and dry grasslands (p-value < 0.001), but with overlapping mean ± SD (Supplementary material Appendix 1 Fig. A8–A9).

Soil property model evaluations

All Random Forest models of the in-situ soil properties (NABO soil database) provided very good test statistics with strong positive correlations when evaluated by split-sample testing (spearman correlation r mean ± SD: NABO-pH = 0.977 ± 0.002, NABO-N = 0.962 ± 0.007, NABO-C = 0.944 ± 0.007, NABO-C:N = 0.968 ± 0.005) and block cross-validation (NABO-pH = 0.981 ± 0.001, NABO-N = 0.957 ± 0.001, NABO-C = 0.949 ± 0.001, NABO-C:N = 0.955 ± 0.001). However, compared to the EIV models, the extrapolation into novel covariate space was much higher (type 1 novelty = ~9.9%; type 2 novelty = ~0.79%; Supplementary material Appendix 1 Fig. A10) and mostly concerned alpine regions or lakes (Supplementary material Appendix 1 Fig. A10) due to the low representation of high elevation sites and lake borders in the calibration data. The generated NABO maps of in-situ soil properties are visualized in Supplementary material Appendix 1 Fig. A11.

EIV importance in SDMs

We found that NDVI mean, NDVI SD, forest height and EIV variables were among the most important predictors in five-variable model tests among 500 plant species. EIVs outperformed traditional climate predictors such as temperature or precipitation, as well as traditional proxies of moisture gradients (Swb, Mind and Water) and soil pH (bedrock geology), and were also more informative than in-situ soil properties modelled from the NABO soil database (Fig. 4). The importance of the EIV predictors for explaining the distribution of the species varied among ecological groups (Fig. 4). EIV-R was one of the most important predictors among all ecological groups (except for ruderal and nutrient-rich grassland plants) and constantly outperformed bedrock geology (Geology). Similarly, EIV-K showed high contributions in explaining plant distributions of almost all ecological groups, but especially of dry grassland plants. EIV-F was particularly important for explaining the distribution of dry grassland and ruderal plants, and strongly outperformed climate and land cover predictors reflecting moisture gradients (Prec, Swb, Mind and Water; Fig. 4). EIV-W and EIV-D were important for explaining the distribution of species growing in moist grassland habitats. EIV-N was important for ruderal and dry grassland plants, and EIV-L was important for plants growing in forests and nutrient-rich grasslands. EIV-H showed in general poor contribution for explaining plant distributions, but was more important for mountain and pioneer plants. Only few in-situ soil properties modelled from the NABO soil database showed better performances than EIVs. NABO-C:N was particularly important for explaining the distribution of plants growing in nutrient-rich grasslands and outperformed EIV-N. In contrast, NABO-pH was constantly outperformed by EIV-R. Beyond the aforementioned predictors, TRI was important for explaining the distribution of mountain plants, and NDVI variation (NDVI SD), a metric reflecting biomass related landuse changes, was particularly important for explaining the distribution of pioneer and ruderal plants.

Species distribution models with and without EIV predictors

Among the 60 plant species tested, SDMs calibrated without EIV predictors showed good predictive performance (TSS = 0.782 ± 0.083; sensitivity = 0.927 ± 0.037; specificity = 0.854 ± 0.061; Supplementary material Appendix 1 Table A3), and so did models including EIV predictors (TSS = 0.840 ± 0.080; sensitivity = 0.939 ± 0.039; specificity = 0.900 ± 0.058; Supplementary material Appendix 1 Table A3). Inclusion of EIV predictors improved SDM performance for a high proportion of species (90% for TSS, 66.7% for sensitivity and 90% for specificity; see Fig. 5 for a few examples). The performance increased on average by 7.7 ± 6.3, 1.3 ± 3.1 and 5.5 ± 5.1% for TSS, sensitivity and specificity, respectively. All algorithms in the ensemble SDMs showed on average increased performance gains and we found no significant differences between them (anova: df = 4, F-value = 2.388, p-value = 0.051; Supplementary material Appendix 1 Table A5). The average performance gain (delta TSS = 0.058 ± 0.045) was mainly driven by an increase in model specificity (rate of true negatives; delta specificity = 0.046 ± 0.040) rather than model sensitivity (rate of true positives; delta sensitivity = 0.012 ± 0.02; Supplementary material Appendix 1 Fig. A12). Species with affinities for calcareous soils showed higher performance gains through inclusion of EIV predictors than species growing on acidic soils (Supplementary material Appendix 1 Fig. A12). Similarly, species growing under highly variable soil moisture conditions showed higher performance gains than species with affinities for soils with more constant soil moisture

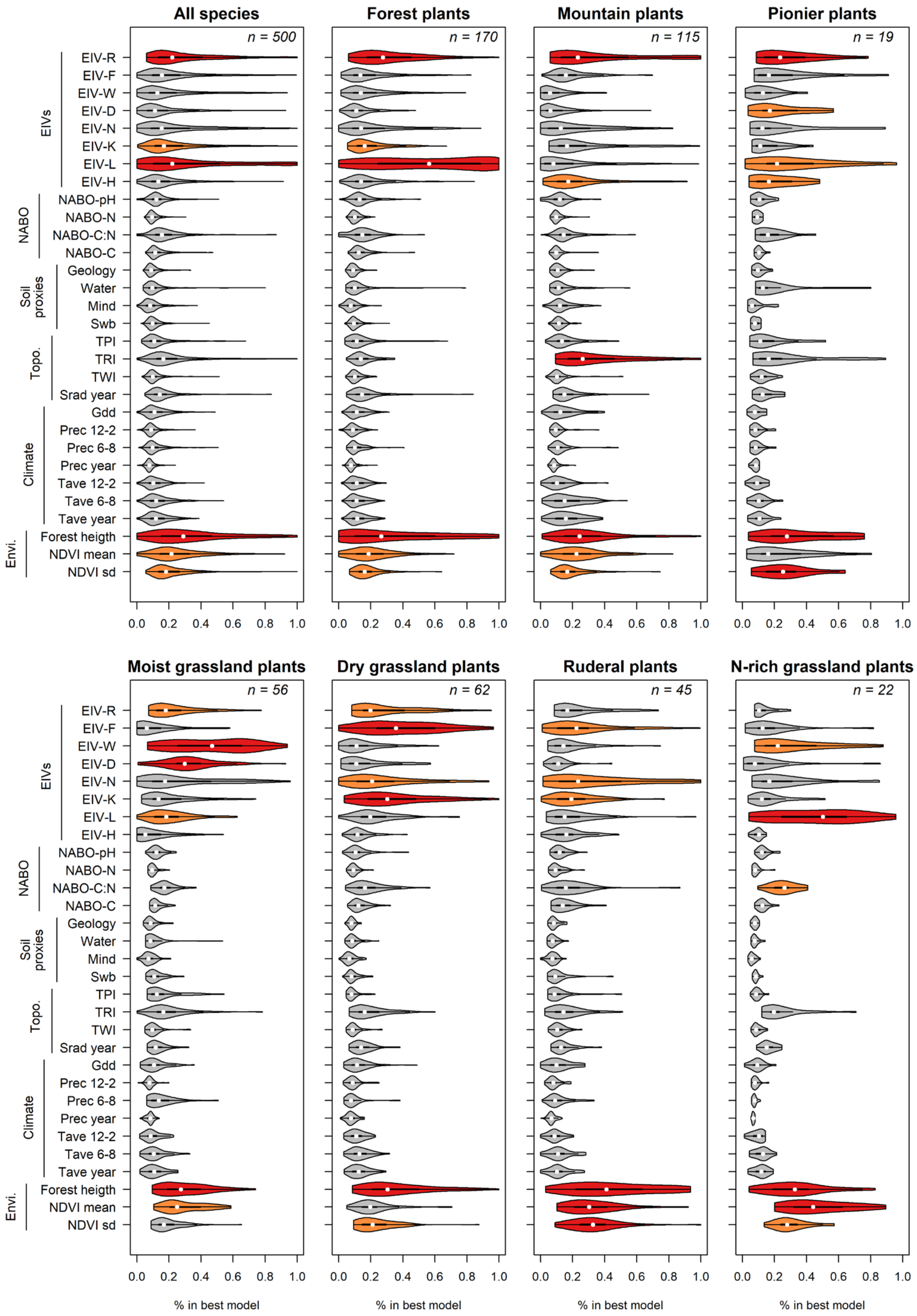


Figure 4. Violin plots depicting the variable importance in different plant ecological groups estimated by testing all possible combinations of five variables with quadratic terms (excluding combinations with collinear variables) with generalized linear models. The variable importance was assessed for 500 species as the percentage of time a variable is present in the 5% best models ranked by their AICs (4042 out of 80 845 models for every species). The three best variables (number one to three) based on their median values (white dot) are highlighted in red, and the next three best variables (number four to six) are highlighted in orange. See Supplementary material Appendix 1 Table A1, Note A2 for a description of the predictors. N-rich = nutrient-rich, Topo. = topography; Envi. = environment.

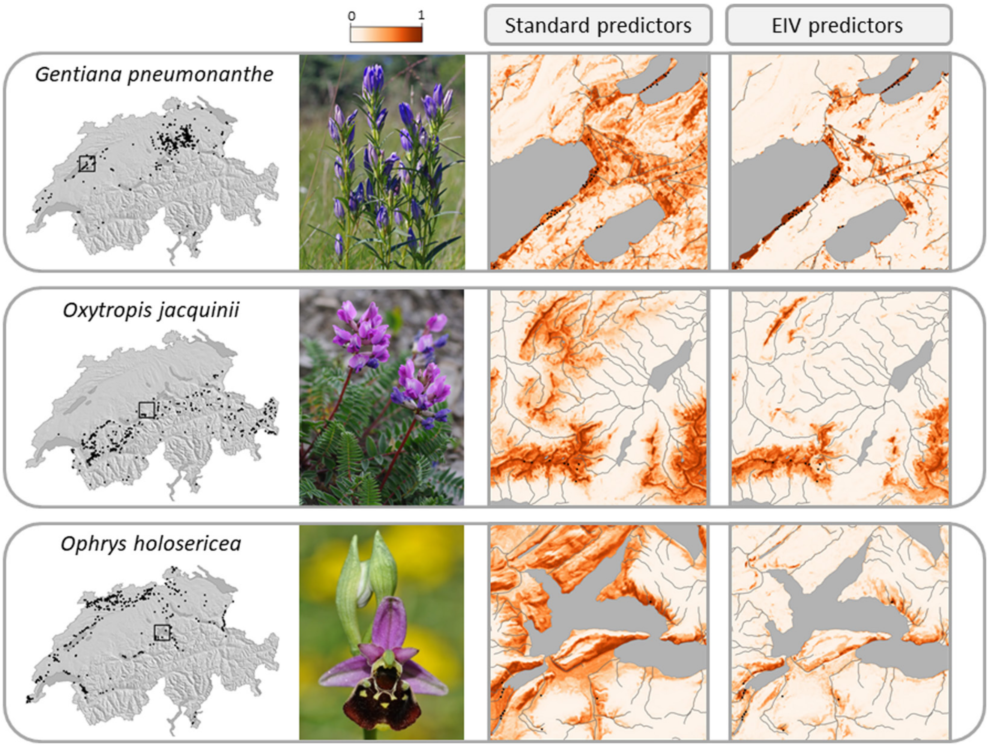


Figure 5. Habitat suitability maps predicted from ensemble SDMs for three of the 60 plant species modelled across Switzerland. Maps illustrate different focal regions (squares of 30 km width) across the study area. SDMs excluded (Standard predictors) or included EIVs predictors (EIVs predictors), respectively, for building models. The Ensemble SDM was calculated by averaging all replicates and algorithms. Species occurrences are represented by black dots in the hillshaded maps on the left and in the focal regions. Lakes and rivers are represented in grey. The scale (0–1) represents the habitat suitability of the species. Photo credits: Joëlle Magnin-Gonze (*G. pneumonanthe* L., *O. jacquinii* Bunge) & Patrice Descombes (*O. holosericeae* (Burm. f.) Greuter).

(Supplementary material Appendix 1 Fig. A12). Also, species growing at the extremes of the soil moisture gradient showed higher gains than species preferring intermediate levels (Supplementary material Appendix 1 Fig. A12).

Discussion

We show that combining large occurrence databases with expert knowledge of plant EIVs is a powerful way for generating ecophysiologically meaningful predictors of local edaphic and climatic conditions at a high spatial resolution (93 m) over complex terrain. EIV models showed very high performances and their predictions were correlated to in-situ physico-chemical properties assessed in soil profiles, emphasizing the relevance of EIVs in reflecting ecological and physico-chemical properties. When used in SDMs, EIV variables outperformed mapped in-situ soil properties and some of the commonly used variables (Fig. 4), and improved the performance of SDMs by 7.7% on average by increasing model specificity and reducing over-predictions of species distributions (Fig. 5, Supplementary material Appendix 1 Fig. A12). Together, our results suggest that combining large occurrence databases with expert knowledge of plant EIVs is a powerful approach for generating additional, ecologically relevant predictors of plant species' distributions.

Relationship between site averaged EIVs and in-situ soil properties

In this study, we first investigated the relationship between in-situ soil properties and site averaged EIVs using exhaustive plant inventories from the WSL and NABO soil databases. Among the EIVs investigated, EIV-R best correlated to in-situ topsoil properties, showing a positive correlation to soil pH. While strong correlations between EIV-R and soil pH are generally reported (Ertsen et al. 1998, Schaffers and Sýkora 2000, Wamelink et al. 2002, Diekmann 2003), EIV-R has also been shown to better reflect the total amount of calcium (exchangeable Ca^{2+} and Ca from carbonates) rather than soil reaction per se (Schaffers and Sýkora 2000), which is a pattern that we could not confirm in our analyses on forest plots (lower correlation of EIV-R with Ca than pH). The strong relationship observed between in-situ soil pH and averaged EIV-R suggests that this EIV is an excellent surrogate of pH measurements.

We found that EIV-N, which mainly stands for nitrogen and phosphorus availability according to Landolt et al. (2010), was negatively correlated to C:N, which corroborates observations by Schaffers and Sýkora (2000). C:N is an indicator of the potential degradability of the organic matter in the soil and therefore reflects the nitrogen availability for plants, where low values indicate higher availability (Andrianarisoa et al. 2009). EIV-N has also been shown to better reflect parameters related to biomass productivity or plant growth rates (e.g. biomass production, leaf nitrogen content; Ertsen et al. 1998, Schaffers and Sýkora 2000) and other soil nutrients (e.g. P, K, NO_3^-, NH_4^+), rather than measurements of soil nitrogen and related parameters (e.g. total nitrogen content, mineralization, C:N; Schaffers and Sýkora 2000, Diekmann 2003). It is therefore not surprising that we found a relatively weak correlation between this EIV and soil measurements such as total nitrogen content or amounts of basic cations.

Soil moisture indicators, such as EIV-F and EIV-D were moderately associated to modelled drought index and soil hydromorphy, respectively. While comparisons between soil moisture EIVs and in-situ measurements are generally scarce in the literature (Diekmann 2003), Ellenberg moisture values (F) have been shown to correlate well with groundwater levels estimated from soil profiles (Ertsen et al. 1998, Schaffers and Sýkora 2000, Wamelink et al. 2002). Similarly, Häring et al. (2013) predicted successfully Ellenberg's soil moisture values in forests of the Bavarian Alps and found that the generated high-resolution map provided valuable surrogates of site hydrological conditions. EIVs offer therefore a valuable alternative to hydrological models predicting soil moisture conditions (Piedallu et al. 2011, Häring et al. 2013, Cianfrani et al. 2019).

Finally, EIV-H was moderately positively associated to soil organic carbon contents in the NABO soil database only. Because the strength of the correlation is dependent on the length of the gradient considered (Diekmann 2003), a higher correlation with data from the NABO database might arise because this database spans a larger number of vegetation types with potentially higher contrasts in soil organic carbon content (range of in-situ C values: 0.1–50.2 $g\,kg^{-1}$), or due to differences in the soil sampling protocol. In comparison, WSL soil inventories were solely performed in forests, where soil organic carbon content might be more evenly distributed, possibly leading to lower contrasts in organic carbon content between the sampled sites and therefore to weaker correlations to EIV-H (range of in-situ C values: 7.6–450.5 $g\,kg^{-1}$).

Overall, our results indicate that EIVs are useful surrogates and descriptors of different soil properties measured in the field. In particular, EIVs show strong potential to augment information gained from site measurements, as they can reflect multiple ecological requirements or conditions which cannot be easily measured in the field at large extents and high spatial resolution (i.e. < 100 m). This strengthens the relevance of predicting EIVs across large geographical areas and their potential use as ecologically meaningful predictors in SDMs.

Mapping EIVs

While spatial interpolation of edaphic and climatic conditions from site measurements is common practice (Schloeder et al. 2001, Piedallu et al. 2011, Pradervand et al. 2014), spatial modelling of site conditions has only recently gained importance (Häring et al. 2013, Buri et al. 2017). However, current challenges in soil mapping have to deal with the generally low amount of soil data available across large geographic extents, which can lead to a high degree of spatial extrapolation into environmental spaces not covered by the calibration data (Mesgaran et al. 2014). This can limit the capacity to accurately predict site characteristics, especially over large spatial extents and complex terrain. In particular, soil profile measurements, which are also generally rare, time consuming and costly to collect and process (Häring et al. 2013, Carter et al. 2015, Mod et al. 2016), might suffer from a low coverage of the environmental space. Hence, even if the NABO soil database in our study covers a large proportion of the extent of Switzerland and includes different types of habitats, we found that the predictions of in-situ soil properties were extrapolated to ca 11% of the area (predictions are outside the predictor range in the calibration data), questioning the robustness of those layers, despite generally good evaluation scores.

Our study overcame these challenges by combining indicator values to large citizen-science plant distribution databases. This enabled us to consider more than 70 000 averaged cell EIV conditions representing an increase of 50× in soil related data information compared to available soil profile samples (NABO: 1131; WSL: 1156) across Switzerland presented in this study. We found that all EIVs were successfully predicted at 93 m resolution across Switzerland. As expected, EIV predictions correlated to some of the measured soil properties, but correlations were generally weaker than with site-averaged EIVs of plant inventories performed on the soil sampling locations. The lower correlation observed for modelled EIV values compared to site-averaged EIVs likely arises because models are calibrated from cells with at least 10 plant species present (and not from complete inventories) and because sites with the measured soil properties were removed to obtain an independent validation of the performance of the models. Because correlations are generally similar, we can conclude that lacks of correlation between modelled EIV values and in-situ measurements most likely result not from inaccurate predictions of EIV but rather from low association between modelled EIV values and specific soil properties. Among the very few studies attempting to spatially map EIVs, Häring et al. (2013) predicted successfully Ellenberg's soil moisture values in forests of the Bavarian Alps and found that the generated high-resolution map provided valuable surrogates of site hydrological conditions. Using spatially mapped EIVs as predictors in SDMs enable better predictions of plant species distributions than other proxy predictors, as they better resolve the fine scale variation in edaphic and climatic properties.

Improving SDMs by using mapped EIVs as predictors

Together with NDVI and forest height, EIV-R, EIV-K and EIV-L were among the most important predictors in SDMs among the 500 species investigated (Fig. 4). When traditional predictors were replaced by more informative soil and climate related EIV predictors in SDMs among the 60 plant species modelled, the performance of SDMs clearly improved. In particular, the performance improvement was mainly driven by an increase in model specificity, meaning that species ranges were less overestimated when including EIV predictors. EIV predictors enabled a finer characterization and delimitation of the species ecological niche and avoided species overprediction to the landscape. For instance, the distribution of *Gentiana pneumonanthe* L. (Fig. 5), a species growing in moist, open habitats, showed more accurate predictions and lower overpredictions when EIV maps where included as SDM predictors. In this case, EIF-F, EIV-W and EIV-D provided much better capacities to discriminate wetlands from mesic and dry grasslands compared to other proxies of soil moisture (distance to water, moisture index and annual average site water balance; Supplementary material Appendix 1 Fig. A7–A9). Furthermore, the SDM performance particularly improved for plants on calcareous soils and plants growing at the extremes of the soil moisture gradient (Supplementary material Appendix 1 Fig. A12). Previous studies already indicated performance gains in plant SDMs when including soil pH (Coudun et al. 2006, Dubuis et al. 2013, Buri et al. 2017) or soil moisture parameters (Cianfrani et al. 2019) as predictors. Scherrer and Guisan (2019) found that ecological conditions reflecting soil properties or characteristics and light availability are, in addition to temperature, very important predictors and should be included in plant distribution modelling. Similarly, in-situ soil properties have been shown to better explain the distribution of tree species in temperate forests than climate variables (Walthert and Meier 2017).

Predictor importance in SDMs varied between species and among the different ecological groups but was overall consistent with species' ecological characteristics. For instance, EIV-W was a strong predictor for explaining the distribution of plant species growing in moist habitats subject to intra-annual water supply variability (e.g. *Gentiana pneumonanthe* L.; Delarze et al. 1998). Similarly, plants growing in forests were best explained by EIV-L, underlining the importance of available light in structured vegetation types with several layers of foliage (Nieto-Lugilde et al. 2015, Scherrer and Guisan 2019). Furthermore, mapped EIVs outperformed some of the traditional predictors commonly used in distribution modelling. For instance, EIV-R was more informative than the reclassified lithologies of the geology map reflecting a gradient of increasing $CaCO_3$, and EIV-W, EIV-F or EIV-D generally outperformed distance to water, moisture index and annual average site water balance parameters and provided much better capacities to discriminate wetlands from dry grasslands. Furthermore, EIVs generally outperformed predictors of in-situ soil properties modelled with the NABO soil database. This suggest that EIV maps are more informative than soil maps derived from in-situ soil properties, which additionally suffer from low data availability and strong spatial extrapolations in model predictions. Surprisingly, temperature parameters were less important for explaining the distribution of plant species, likely because some of the EIVs are already correlated to temperature. Because all models included the same number of variables, the observed improvement in model performance resulted from the additional information expressed by EIV not accounted for by traditional environmental predictors (Scherrer and Guisan 2019).

While biodiversity databases are hosting increasing numbers of species occurrences through citizen science contributions, there is high potential for extending our approach to analyzing larger spatial extents such as all of Europe or the globe. Generating new EIVs or merging EIV databases in the same taxonomic backbone (Dengler et al. 2016), for all or a few key target species (species with affinities for acidic or calcareous soils, or affinities for moist or dry soil conditions), provides a cost-effective alternative to traditional site measurements for modelling spatially explicit edaphic and climatic properties at large spatial scale. Finally, EIVs can also be inferred from soil pit data where plants were also inventoried (Diekmann and Falkengren-Grerup 1998, Wamelink et al. 2002), and then be used in combination with biodiversity databases to model soil properties at large spatial extents.

Limitations

While comprehensive and easy to apply, our approach is based on few important assumptions. First, EIVs represent expert-based and subjective quantifications of the response of species to environmental conditions. The EIV maps in this study are therefore only surrogates of edaphic and climate conditions constrained between a defined range of values and representing contrasts in ecological conditions in the landscape.

Second, site-averaging ordinal values of EIVs is mathematically incorrect. But previous studies have demonstrated that EIVs averaged at the plot level are excellent bioindicators of the site characteristics when plant inventories are available (Ter Braak and Barendregt 1986, Scherrer and Guisan 2019). Furthermore, we did not consider within-species variability in EIV values, nor abundance-based weighting to obtain site mean EIV values. This rather simple approach of averaging site EIVs among jointly occurring species tend to constrain the average EIV toward the mean value of the EIV scales and fails to detect its extremes. Including species EIV variability or abundance-based weighting might enable a better characterisation of extreme site conditions in future analyses and has the potential to reduce the effect of generalist species growing in a large spectrum of ecological conditions towards the mean value of the EIV scale. Similarly, EIV averaging and modelling using higher resolution predictors (e.g. 10 m, 25 m), if available, might provide even finer predictions of EIVs and less blurred EIV estimates in response to subtle changes of the topography.

Third, our approach should primarily be used in areas where information about plant species distributions and their associated EIV are abundantly available. In addition, spatial (i.e. regions with different flora) and temporal (i.e. future) extrapolation should be performed with care. Because EIVs often reflect combinations of factors influencing the growth and performance of the species, predictions of EIVs under future conditions using predictions of atmospheric acidic and nitrogen deposition might sound like a promising, but certainly very complex task. We believe that it might at least be possible to test this for EIV-R, which showed good correlations with in-situ soil pH ($r > 0.73$). But, soil properties might also change due to modified plant species pools under future climate conditions (e.g. nitrogen fixation by symbiotic rhizobia in Fabaceae) or altered soil biotic activity with influences on humus cycling (mineralization) and nutrient availability. Thus, atmospheric acidic and nitrogen deposition will not be the only sources of changes in soil properties, and changes are likely to be multidimensional. Similarly, extrapolation of EIVs into the future might be problematic if biotic factors (such as NDVI or forest height) are included and especially if they show high variable importance in the EIV modelling. In the latter case, biotic factors bring additional levels of circularity as we would either assume that they will change in a certain direction or remain constant, which in fact is just what we want to model based on EIVs. Therefore, further studies are necessary to evaluate the transferability of this approach in space and time.

Finally, because we used plant occurrences to derive EIV maps and then used them to predict the potential distribution of plants, our approach may seem circular. However, we found very similar site averaged EIVs calculated from the occurrence database if single species were iteratively removed from the occurrence database (Spearman correlation: $r > 0.999$; cell loss: $< 0.071\%$). This suggests that our approach of averaging site EIV and using only cells with at least 10 different species enables to robustly capture site conditions and clearly demonstrates that the estimation of the site EIVs are not dependent on single species. Plants can influence local soil and temperature properties (Reich et al. 2005, Aalto et al. 2013). Therefore, site properties are expected to be even more reliable by considering vegetation (Aalto et al. 2013). Our approach provides therefore new insights in generating ecologically meaningful predictors of plant distributions at high spatial resolution (i.e. < 100 m).

Conclusion

The performance of fine-scale SDMs clearly increases if ecologically meaningful predictors are included that better reflect the physiological constraints of species. Together with rapidly growing plant occurrence databases, EIV mapping provides an alternative and cost-effective approach to chart local edaphic and climatic conditions at high spatial resolution (< 100 m) across large geographical scales. The development of ecologically meaningful predictors for plants sets the ground for an increased collaboration between ecological, geo-environmental sciences and biodiversity conservation measures.

Data availability statement

EIV maps and predictors used and generated in this study are made available through EnviDat (<www.envidat.ch>; doi:10.16904/envidat.153). Other datasets and predictors can be accessed via the original data provider.

Acknowledgements – We would like to thank InfoFlora and the Swiss Biodiversity Monitoring programme who kindly provided the data, and Tobias Roth, Michael Jutzi and Daniel Wächter for help with data extraction. We also thank two anonymous referees who provided valuable comments to improve the manuscript.

Funding – This study was supported by the Swiss Data Science Center (SDSC) grant no. c17-07 (SPEEDMIND) project to NEZ, DZ and DNK.

Author contributions – PD designed the study with the help of NEZ, DZ and DNK; PD, LW, AB, RGM and CG collected and prepared the data; P.D. conducted the analyses; and P.D. led the writing of the manuscript with the help of all co-authors.

References

Aalto, J. et al. 2013. Vegetation mediates soil temperature and moisture in arctic–alpine environments. – Arct. Antarct. Alp. Res. 45: 429–439.

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Andrianarisoa, K. S. et al. 2009. Comparing indicators of N status of 50 beech stands (*Fagus sylvatica* L.) in northeastern France. – For. Ecol. Manage. 257: 2241–2253.

Austin, M. P. and Meyers, J. A. 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. – For. Ecol. Manage. 85: 95–106.

Barton, K. 2019. MuMIn: multi-model inference. R package ver. 1.43.6. – <https://CRAN.R-project.org/package=MuMIn>.

Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.

Buri, A. et al. 2017. Soil factors improve predictions of plant species distribution in a mountain environment. – Progr. Phys. Geogr. Earth Environ. 41: 703–722.

Buri, A. et al. 2020. What are the most crucial soil variables for predicting the distribution of mountain plant species? A comprehensive study in the Swiss Alps. – J. Biogeogr. 47: 1143–1153.

Carter, A. L. et al. 2015. Modelling the soil microclimate: does the spatial or temporal resolution of input parameters matter? – Front. Biogeogr. 7: 4.

Cianfrani, C. et al. 2019. Spatial modelling of soil water holding capacity improves models of plant distributions in mountain landscapes. – Plant Soil 438: 57–70.

Coudun, C. and Gégout, J.-C. 2007. Quantitative prediction of the distribution and abundance of *Vaccinium myrtillus* with climatic and edaphic factors. – J. Veg. Sci. 18: 517–524.

Coudun, C. et al. 2006. Soil nutritional factors improve models of plant species distribution: an illustration with *Acer campestre* (L.) in France. – J. Biogeogr. 33: 1750–1763.

Delarze, R. et al. 1998. Guide des milieux naturels de Suisse: écologie, menaces, espèces caractéristiques. – Delachaux et niestlé.

Dengler, J. et al. 2016. Ecological indicator values of Europe (EIVE) 1.0: a powerful open-access tool for vegetation scientists. – Oral presentation and abstract of the 25th European Vegetation Survey Meeting 6–9 April 2016 in Rome, IT.

Diekmann, M. 1995. Use and improvement of Ellenberg's indicator values in deciduous forests of the Boreo-nemoral zone in Sweden. – Ecography 18: 178–189.

Diekmann, M. 2003. Species indicator values as an important tool in applied plant ecology – a review. – Basic Appl. Ecol. 4: 493–506.

Diekmann, M. and Falkengren-Grerup, U. 1998. A new species index for forest vascular plants: development of functional indices based on mineralization rates of various forms of soil nitrogen. – J. Ecol. 86: 269–283.

Dormann, C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 36: 27–46.

Dubuis, A. et al. 2013. Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. – J. Veg. Sci. 24: 593–606.

Ellenberg, H. et al. 1992. Zeigerwerte von Pflanzen in Mitteleuropa. – Scripta Geobot. 18: 1–248.

Elmendorf, S. C. and Moore, K. A. 2008. Use of community-composition data to predict the fecundity and abundance of species. – Conserv. Biol. 22: 1523–1532.

Ertsen, A. C. D. et al. 1998. Calibrating Ellenberg indicator values for moisture, acidity, nutrient availability and salinity in the Netherlands. – Plant Ecol. 135: 113–124.

Figueiredo, F. O. G. et al. 2018. Beyond climate control on species range: the importance of soil data to predict distribution of Amazonian plant species. – J. Biogeogr. 45: 190–200.

Fowler, J. et al. 2013. Practical statistics for field biology. – Wiley.

Gégout, J.-C. et al. 2003. Prediction of forest soil nutrient status using vegetation. – J. Veg. Sci. 14: 55–62.

Grunwald, S. et al. 2011. Digital soil mapping and modeling at continental scales: finding solutions for global issues. – Soil Sci. Soc. Am. J. 75: 1201–1213.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.

Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.

Guisan, A. et al. 2017. Habitat suitability and distribution models: with applications in R. – Cambridge Univ. Press.

Häring, T. et al. 2013. Predicting Ellenberg's soil moisture indicator value in the Bavarian Alps using additive georegression. – Appl. Veg. Sci. 16: 110–121.

Hengl, T. et al. 2017. SoilGrids250m: global gridded soil information based on machine learning. – PLoS One 12: e0169748.

Heung, B. et al. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. – Geoderma 265: 62–77.

Körner, C. 2003. Alpine plant life: functional plant ecology of high mountain ecosystems; with 47 tables. – Springer.

Landolt, E. et al. 2010. Flora indicativa: Ökologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen. – Haupt.

Lenoir, J. et al. 2013. Local temperatures inferred from plant communities suggest strong spatial buffering of climate warming across Northern Europe. – Global Change Biol. 19: 1470–1481.

McCullagh, P. 1983. Generalized linear models. – Eur. J. Operat. Res. 16: 285–292.

Mesgaran, M. B. et al. 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. – Divers. Distrib. 20: 1147–1159.

Meuli, R. G. et al. 2017. Connecting biodiversity monitoring with soil inventory information – a Swiss case study. – BGS Bull. 38: 65–69.

Mod, H. K. et al. 2016. What we use is not what we know: environmental predictors in plant distribution models. – J. Veg. Sci. 27: 1308–1322.

Nieto-Lugilde, D. et al. 2015. Tree cover at fine and coarse spatial grains interacts with shade tolerance to shape plant species distributions across the Alps. – Ecography 38: 578–589.

Nussbaum, M. et al. 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. – Soil 4: 1–22.

Padarian, J. et al. 2019. Using deep learning for digital soil mapping. – Soil 5: 79–89.

Pebesma, E. J. 2004. Multivariable geostatistics in S: the gstat package. – Comput. Geosci. 30: 683–691.

Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.

Piedallu, C. et al. 2011. Mapping soil water holding capacity over large areas to predict potential production of forest stands. – Geoderma 160: 355–366.

Pinto, P. E. and Gégout, J.-C. 2005. Assessing the nutritional and climatic response of temperate tree species in the Vosges Mountains. – Ann. For. Sci. 62: 761–770.

Pradervand, J.-N. et al. 2014. Very high resolution environmental predictors in species distribution models: moving beyond topography? – Progr. Phys. Geogr. Earth Environ. 38: 79–96.

Reich, P. B. et al. 2005. Linking litter calcium, earthworms and soil properties: a common garden test with 14 tree species. – Ecol. Lett. 8: 811–818.

Roberts, D. R. et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. – Ecography 40: 913–929.

Robinson, N. et al. 2014. EarthEnv-DEM90: a nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. – ISPRS J. Photogrammetry Remote Sens. 87: 57–67.

Rydin, H. et al. 2006. The role of sphagnum in Peatland development and persistence. – In: Wieder, R. K. and Vitt, D. H. (eds), Boreal peatland ecosystems. Ecological studies. Springer, pp. 47–65.

Schaffers, A. P. and Sýkora, K. V. 2000. Reliability of Ellenberg indicator values for moisture, nitrogen and soil reaction: a comparison with field measurements. – J. Veg. Sci. 11: 225–244.

Scherrer, D. and Guisan, A. 2019. Ecological indicator values reveal missing predictors of species distributions. – Sci. Rep. 9: 3061.

Schloeder, C. A. et al. 2001. Comparison of methods for interpolating soil properties using limited data. – Soil Sci. Soc. Am. J. 65: 470–479.

Smart, S. M. et al. 2003. National-scale vegetation change across Britain; an analysis of sample-based surveillance data from the countryside surveys of 1990 and 1998. – J. Environ. Manage. 67: 239–254.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. – Ecol. Lett. 10: 1115–1123.

Strobl, C. et al. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. – BMC Bioinform. 8: 25.

Ter Braak, C. J. F. and Barendregt, L. G. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. – Math. Biosci. 78: 57–72.

Thuiller, W. 2013. On the importance of edaphic variables to predict plant species distributions – limits and prospects. – J. Veg. Sci. 24: 591–592.

Walthert, L. and Meier, E. S. 2017. Tree species distribution in temperate forests is more influenced by soil than by climate. – Ecol. Evol. 7: 9473–9484.

Walthert, L. et al. 2013. Shortage of nutrients and excess of toxic elements in soils limit the distribution of soil-sensitive tree species in temperate forests. – For. Ecol. Manage. 297: 94–107.

Wamelink, G. W. W. et al. 2002. Validity of Ellenberg indicator values judged from physico-chemical field measurements. – J. Veg. Sci. 13: 269–278.

Wamelink, G. W. W. et al. 2005. Plant species as predictors of soil pH: replacing expert judgement with measurements. – J. Veg. Sci. 16: 461–470.

Wohlgemuth, T. et al. 1999. Computed ecograms of Swiss forests. – Bot. Helv. 109: 169–191.

Woodward, F. I. 1987. Climate and plant distribution. – Cambridge Univ. Press.

Zellweger, F. et al. 2020. Forest microclimate dynamics drive plant responses to warming. – Science 368: 772–775.

Zimmermann, N. E. and Kienast, F. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. – J. Veg. Sci. 10: 469–482.

Supplementary material (available online as Appendix ecog-05117 at <www.ecography.org/appendix/ecog-05117>). Appendix 1.